STAC67 A2

Feifei Fu 1006740216 Wenqing Liang 1006739709

2022-10-19

Question1

a)

```
data <- read.table("vote-1.txt", header = TRUE)
set.seed(1006740216)

## step1
n = nrow(data)
x = data$growth
e = rnorm(n,0,3.9)
y = 46.3 + 4*x + e

## step2
lm = lm(y~x)
```

```
summary(lm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5024 -2.3540  0.0432  3.3270  5.5689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.0380     1.8780  23.449 1.23e-12 ***
## x             4.0651     0.8062   5.042  0.00018 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.358 on 14 degrees of freedom
## Multiple R-squared:  0.6449, Adjusted R-squared:  0.6195
## F-statistic: 25.42 on 1 and 14 DF,  p-value: 0.0001799
```

```
mean(x)
```

```
## [1] 1.8975
```

$\hat{\beta}_0 = 44.0380$

$\hat{\beta}_1 = 4.0651$

By hand, 95% CI = ($\hat{y}$ - $t_{0.975,n-2} * se(\hat{y})$, $\hat{y} + t_{0.975,n-2} * se(\hat{y})$)

$\hat{y}$ = 44.0380 + 0.1 * 4.0651 = 44.44451

$\hat{\sigma}$ = 4.358

```
qt(0.975, n-2)
```

```
## [1] 2.144787
```

```
44.44451 + qt(0.975, n-2) * sqrt((1/n + ((0.1 - mean(x))^2 / sum((x - mean(x))^2))) * 4.358^2 )
```

```
## [1] 48.33337
```

```
44.44451 - qt(0.975, n-2) * sqrt((1/n + ((0.1 - mean(x))^2 / sum((x - mean(x))^2))) * 4.358^2 )
```

```
## [1] 40.55565
```

Therefore, 95% CI = (40.55565, 48.33337)

```
predict(lm, data.frame(x=0.1) ,interval = "confidence")
```

```
##       fit      lwr      upr
## 1 44.4445 40.55606 48.33294
```

By R, 95% CI = (40.55606, 48.33294) is close to the result by hand.

b)

```
set.seed(1006740216)
beta0 <- c()
beta1 <- c()

for (i in 1:10000){

 x = data$growth
 e = rnorm(n,0,3.9)
 y = 46.3 + 4*x + e
 lm = lm(y~x)

beta0[i] = summary(lm)$coefficients[1]
beta1[i] = summary(lm)$coefficients[2]
}
```

The histogram of $\beta 0$:

```
mean0 = 46.3
sd0 = sqrt(3.9^2 * (1/n+(mean(x))^2/sum((x-mean(x))^2)))
sd0
```

```
## [1] 1.680853
```

```
meannew = mean(beta0)
meannew
```

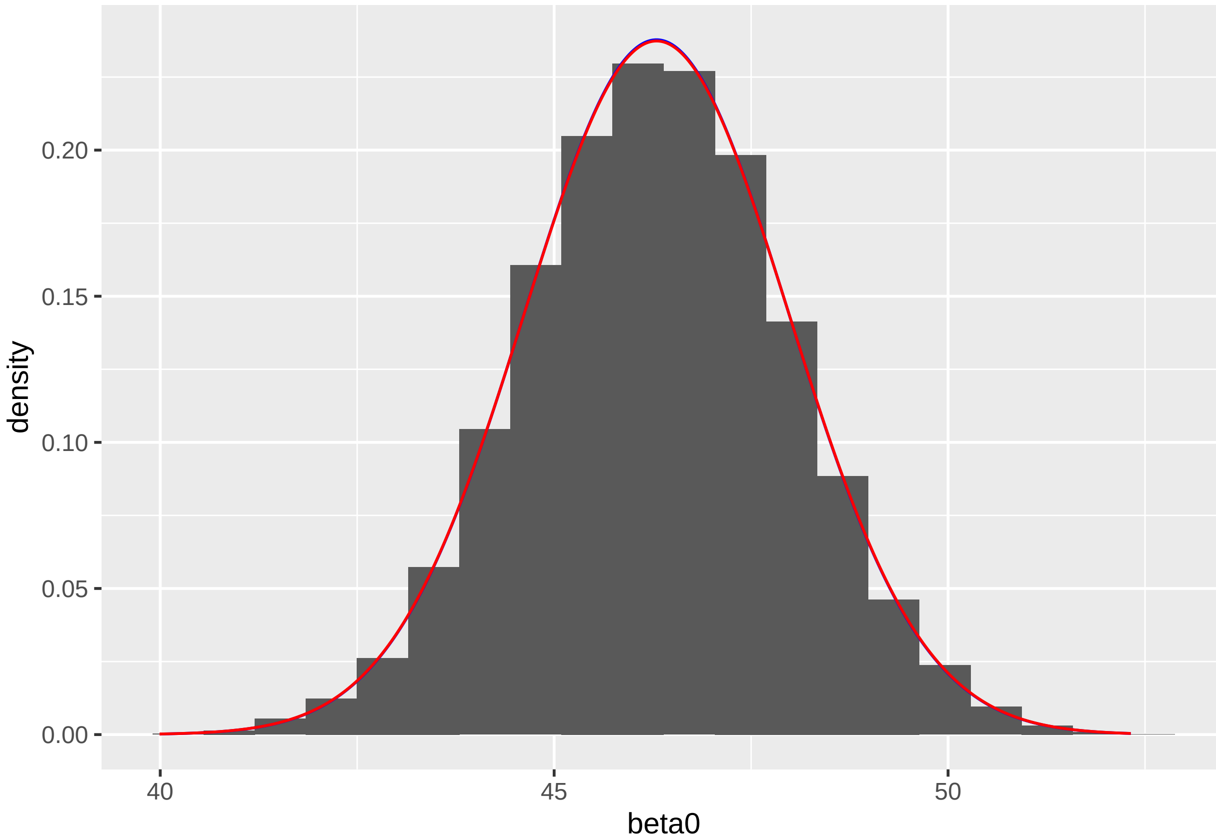
```
## [1] 46.29947
```

```r
sdnew = sd(beta0)
sdnew
```

```
## [1] 1.677881
```

mean of the theoretical distribution = $\beta0 = 46.3$ sd of the theoretical distribution = 1.680853 mean of 10,000 estimates = 46.29947 standard deviation of 10,000 estimates = 1.677881

```r
ggplot(data.frame(beta0), aes(beta0)) + geom_histogram(bins=20, aes(y=..density..)) + stat_function(fun
```



For $\beta0$, for the plot and the number is very close, we can see the results are consistent with theoretical values.

The histogram of $\beta1$:

```r
mean1 = 4
sd1 = sqrt(3.9^2/sum((x-mean(x))^2))
sd1
```

```
## [1] 0.721568
```

```r
meannew1 = mean(beta1)
meannew1
```

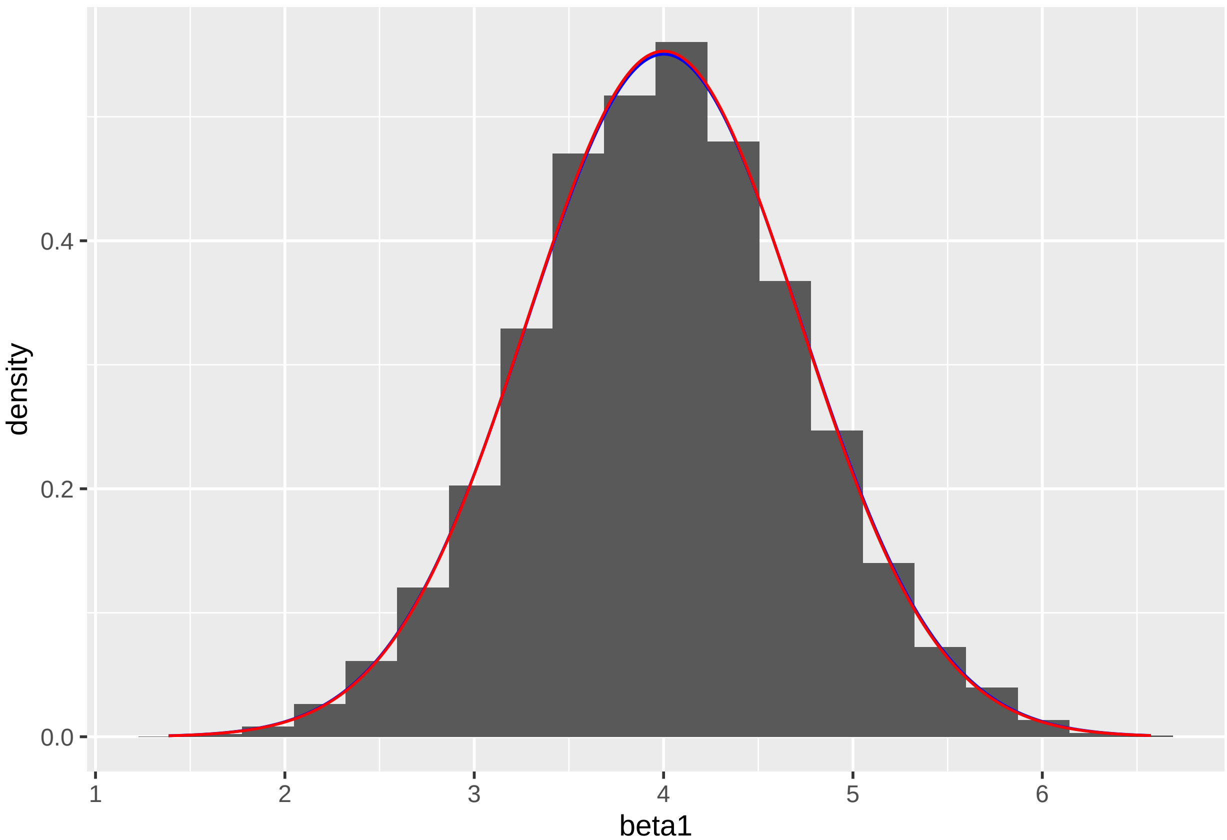
```
## [1] 4.00176
```

```r
sdnew1 = sd(beta1)
sdnew1
```

```
## [1] 0.7247481
```

mean of the theoretical distribution = $\beta1 = 4$ sd of the theoretical distribution = 0.721568 mean of 10,000 estimates = 4.00176 standard deviation of 10,000 estimates = 0.7247481

```
ggplot(data.frame(beta1), aes(beta1)) + geom_histogram(bins=20, aes(y=..density..))  + stat_function(fu
```



For $\beta 1$, for the plot and the number is very close, we can see the results are consistent with theoretical values.

Therefore, the mean and standard deviation of 10,000 estimates $\beta 1$ and $\beta 0$ is consistent with theoretical values.

c)

E(Y|X=0.1) = 46.3 + 4*0.1 = 46.7

```
num = 0
eyx = 46.7

for (i in 1:10000){

 x = data$growth
 e = rnorm(n, 0, 3.9)
 y = 46.3 + 4*x + e
 lm = lm(y~x)

 conf1 = predict(lm, data.frame(x=0.1) ,interval = "confidence")

 low = conf1[2]
 high = conf1[3]

 if (eyx<high & eyx>low) {
   num = num + 1
```

```
  }

}

num/10000
```

```
## [1] 0.9475
```

Proportion of the 10,000 confidence intervals for E(Y|X=0.1) includes E(Y|X=0.1) is: 0.9517

The 95% CI means that 95% of the confidence intervals will contain the value of E(Y|X=0.1).

0.9517 is very close to the 95%. So this result is consistent with theoretical expressions.

Question2

a)

```
data2 <- read.csv("NBAhtwt.csv")
head(data2)
```

```
##              Player Pos Height Weight Age
## 1  Nate\xa0Robinson   G     69    180  29
## 2  Isaiah\xa0Thomas   G     69    185  24
## 3   Phil\xa0Pressey   G     71    175  22
## 4   Shane\xa0Larkin   G     71    176  20
## 5      Ty\xa0Lawson   G     71    195  25
## 6 John\xa0Lucas III   G     71    157  30
```

```
n = 505
x = data2$Height
y = data2$Weight
lm2 = lm(y~x)

summary(lm2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.583  -9.937  -0.260   9.417  56.079
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -279.8693    15.5512  -18.00   <2e-16 ***
## x              6.3307     0.1965   32.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.24 on 503 degrees of freedom
## Multiple R-squared:  0.6736, Adjusted R-squared:  0.6729
## F-statistic:  1038 on 1 and 503 DF,  p-value: < 2.2e-16
```

By hand, 95% CI = ($\hat{y}$ - $t_{0.975,n-2} * se(\hat{y})$, $\hat{y} + t_{0.975,n-2} * se(\hat{y})$)

```
-279.8693 + 6.3307 * 74
```

```
## [1] 188.6025
```

$\hat{y}$ = -279.8693 + 74 * 6.3307 = 188.6025

```
qt(0.975, n-2)
```

```
## [1] 1.964691
```

```
188.6025 + qt(0.975, 503) * sqrt((1/n + (74 - mean(x))^2 / sum((x - mean(x))^2)) * 15.24^2 )
```

```
## [1] 190.9691
```

```
188.6025 - qt(0.975, 503) * sqrt((1/n + (74 - mean(x))^2 / sum((x - mean(x))^2)) * 15.24^2 )
```

```
## [1] 186.2359
```

Therefore, by hands, 95% CI is (186.2359, 190.9691)

By R:

```
n = 505
x = data2$Height
y = data2$Weight
lm2 = lm(y~x)
predict(lm2, data.frame(x=74) ,interval = "confidence")
```

```
##        fit      lwr     upr
## 1 188.6058 186.2397 190.972
```

Therefore, by R, 95% CI is (186.2397, 190.972) is close to the result by hand.

b)

$\hat{y}$ = -279.8693 + 74 * 6.3307 = 188.6025

```
qt(0.975, n-2)
```

```
## [1] 1.964691
```

```
188.6025 + qt(0.975, n-2) * sqrt((1+ 1/n + (74 - mean(x))^2 / sum((x - mean(x))^2)) * 15.24^2 )
```

```
## [1] 218.6378
```

```
188.6025 - qt(0.975, n-2) * sqrt((1+ 1/n + (74 - mean(x))^2 / sum((x - mean(x))^2)) * 15.24^2 )
```

```
## [1] 158.5672
```

Therefore, by hand, 95% prediction interval for a new player with X_0 = 74 is (158.5672, 218.6378).

By R,

```
predict(lm2, data.frame(x=74) ,interval = "predict")
```

```
##        fit      lwr      upr
## 1 188.6058 158.5761 218.6356
```

Therefore, by hand, 95% prediction interval for a new player with X_0 = 74 is (158.5761, 218.6356).

c)

```r
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x           1 240985  240985    1038 < 2.2e-16 ***
## Residuals 503 116782     232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova table is above.

```r
240985/(240985 + 116782)
```

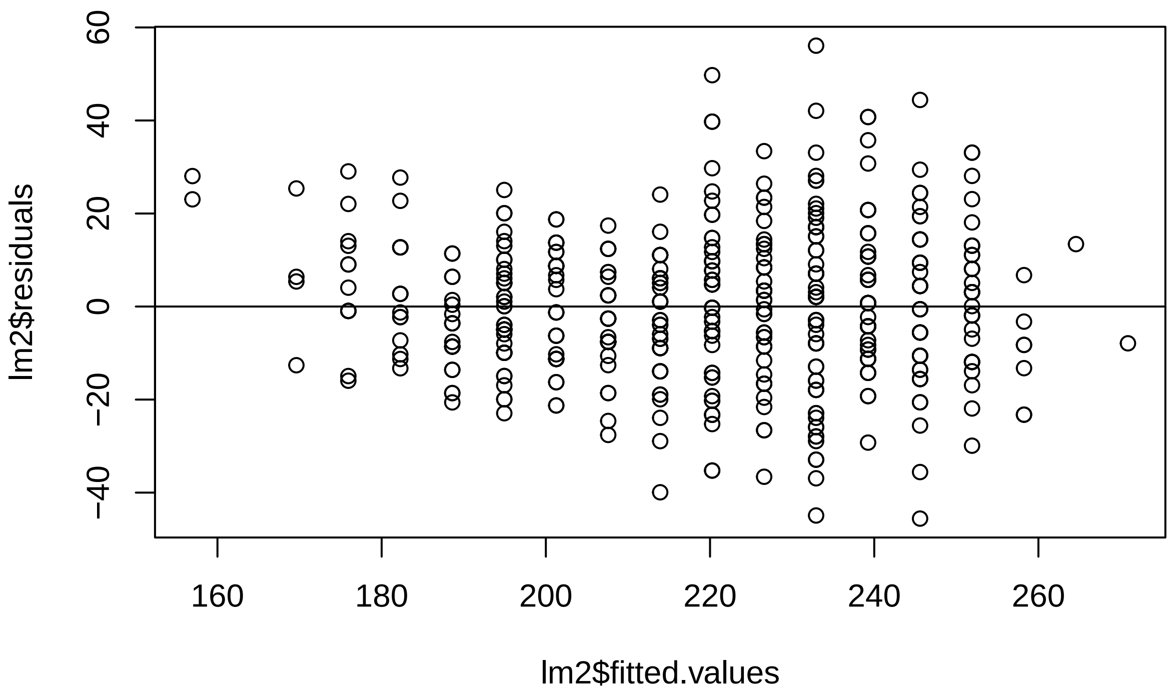
```
## [1] 0.6735809
```

R^2 **= SSR/SST = SSR/(SSR+SSE) = 240985/(240985 + 116782) = 0.6735809**

R^2**: 67.35809% variation in weight can be explained by height.**

d)

```r
plot(lm2$fitted.values, lm2$residuals)
abline(c(0,0))
```

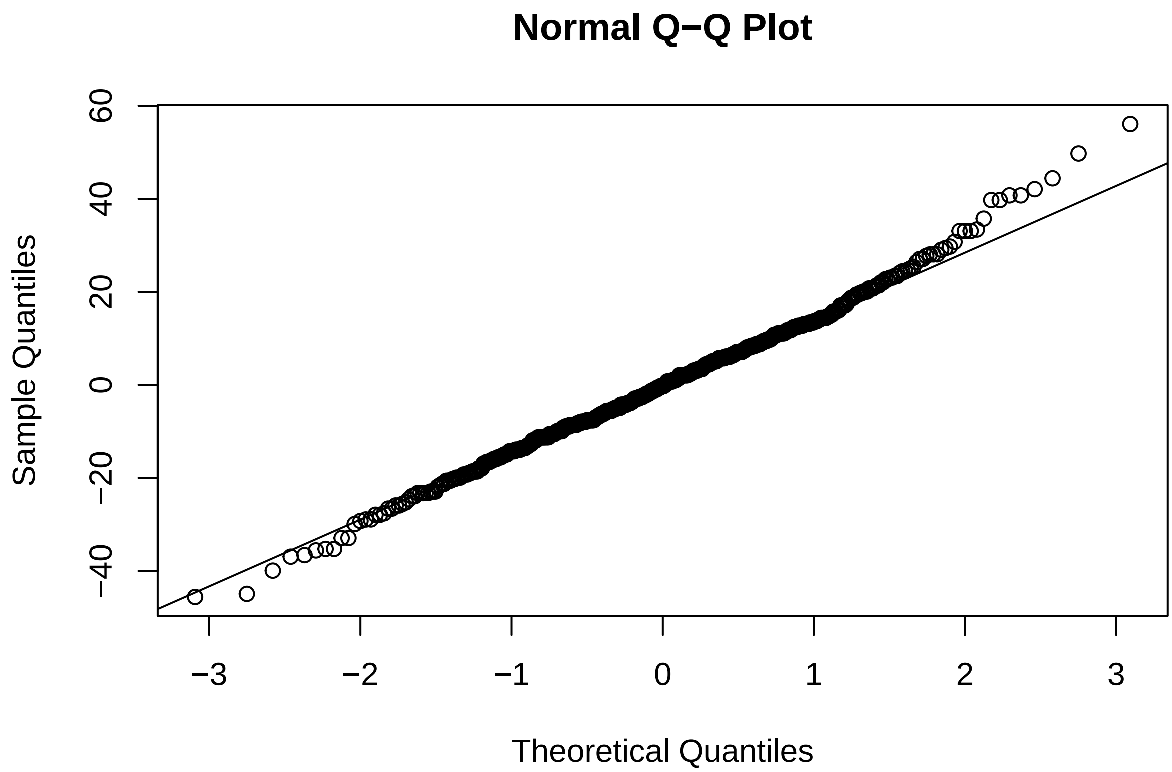


##

The plot looks random but the variance might not be constant

e).

```
qqnorm(lm2$residuals)
qqline(lm2$residuals)
```



H_0: the errors are normally distributed H_1: the errors are not normally distributed

```
shapiro.test(lm2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm2$residuals
## W = 0.9948, p-value = 0.08593
```

Since the p-value is equal to 0.08593, which is bigger than 0.5, we don't have sufficient evidence to reject H_0. Fail to reject H_0, so errors are normally distributed.

f).

```
data3 <- data2 %>% mutate(belowmedian = Height<median(Height))
data3_1 <- data3 %>% filter(belowmedian)
data3_2 <- data3 %>% filter(!belowmedian)
glimpse(data3_1)
```

```
## Rows: 252
## Columns: 6
## $ Player      <chr> "Nate\xa0Robinson", "Isaiah\xa0Thomas", "Phil\xa0Pressey",~
## $ Pos         <chr> "G", "G", "G", "G", "G", "G", "G", "G", "G", "G", "G", "G"~
## $ Height      <int> 69, 69, 71, 71, 71, 71, 72, 72, 72, 72, 72, 72, 72, 72, 72~
## $ Weight      <int> 180, 185, 175, 176, 195, 157, 180, 205, 175, 185, 190, 189~
## $ Age         <int> 29, 24, 22, 20, 25, 30, 25, 27, 28, 30, 31, 24, 28, 22, 25~
## $ belowmedian <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE~
```

```
glimpse(data3_2)
```

```
## Rows: 253
## Columns: 6
## $ Player      <chr> "Brandon\xa0Bass", "Reggie\xa0Evans", "Mirza\xa0Teletovic"~
## $ Pos         <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F"~
## $ Height      <int> 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80~
## $ Weight      <int> 250, 245, 235, 230, 260, 220, 218, 248, 230, 235, 240, 230~
## $ Age         <int> 28, 33, 28, 29, 22, 25, 24, 28, 27, 33, 20, 25, 23, 30, 28~
## $ belowmedian <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA~
```

```
n1 <- nrow(data3_1)
n2 <- nrow(data3_2)
lm3_1 <- lm(Weight ~ Height, data3_1)
lm3_2 <- lm(Weight ~ Height, data3_2)
di1 <- abs(lm3_1$residuals - median(lm3_1$residuals))
di2 <- abs(lm3_2$residuals - median(lm3_2$residuals))
d1 <- mean(di1)
d2 <- mean(di2)
s1_sq <- var(di1)
s2_sq <- var(di2)
var <- ((n1-1)*s1_sq + (n2-1)*s2_sq) / (n1 + n2 -2)
(d1 - d2) / sqrt(var*(1/n1 + 1/n2))
```

```
## [1] -2.519727
```

```
qt(0.975, n1 + n2 -2)
```

```
## [1] 1.964691
```

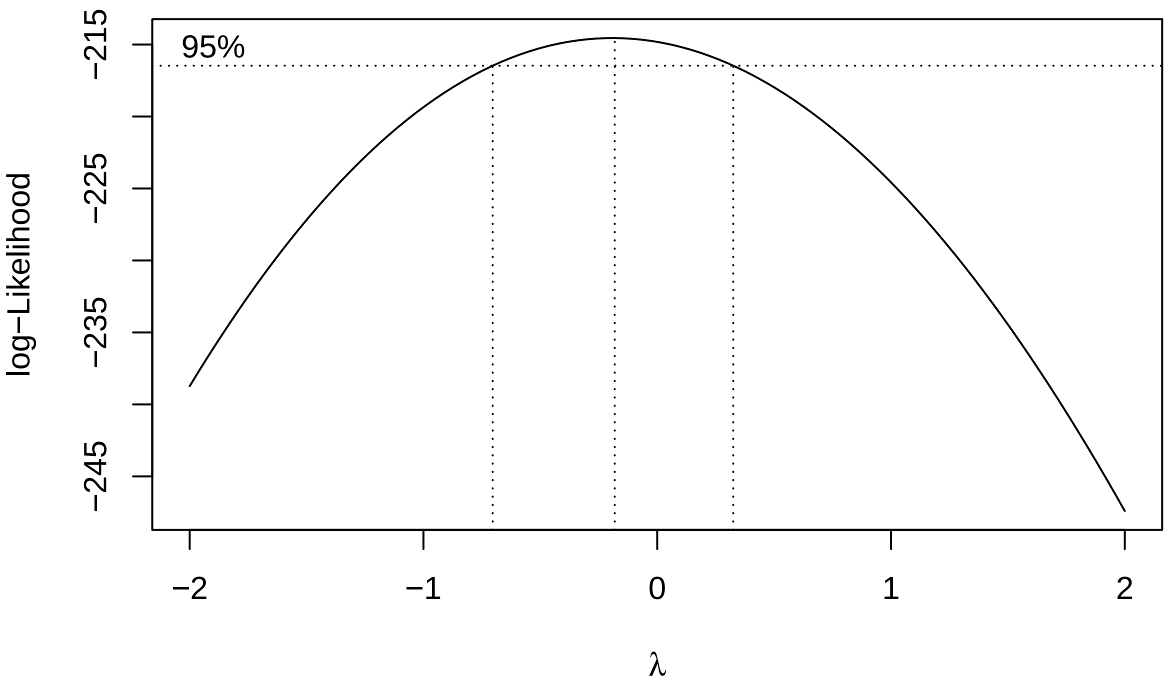
Since $|t_{BF}| = 2.519727$, which is bigger than $t_{0.975,n-2} = 1.964691$, we have enough evidence to reject H_0. Thus, the error variance varies.

g).

```
library(MASS)
```

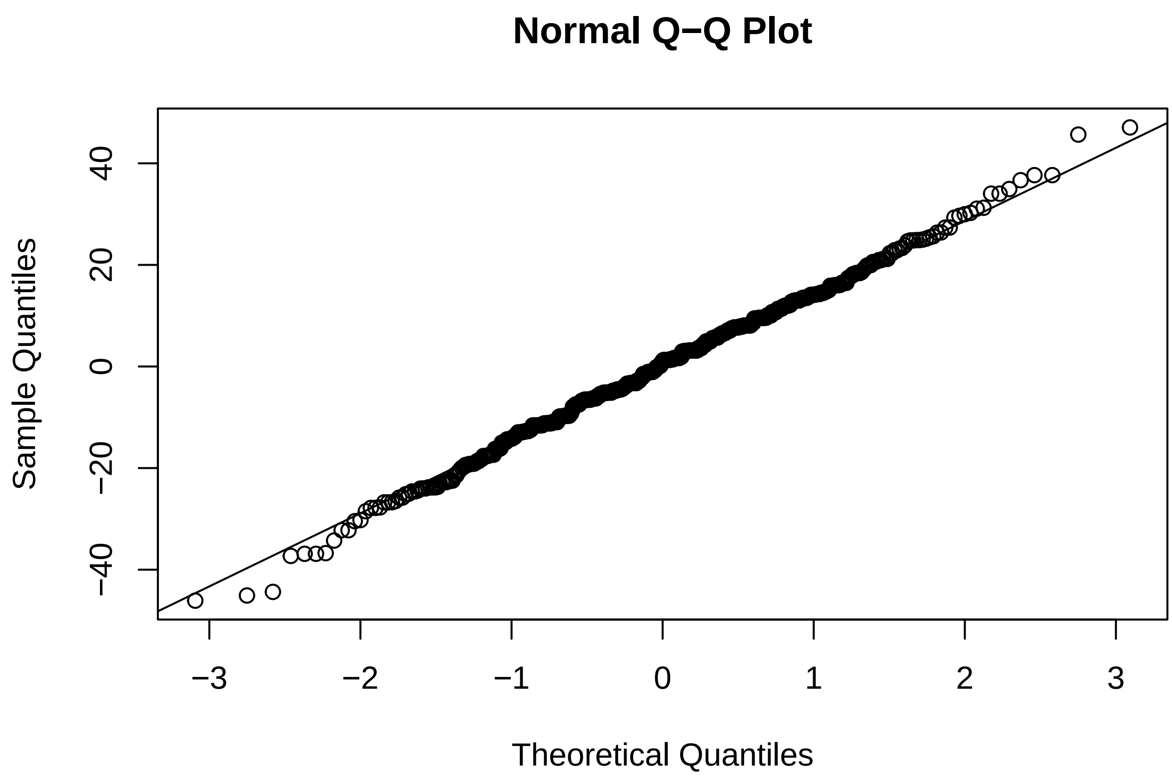
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
result = boxcox(lm)
```




```
lambda = result$x[which.max(result$y)]
X = data2$Height
Y = data2$Weight
k2 = exp(sum(log(y)) / n)
k1 = 1 / (lambda * k2^(lambda - 1))
w = k1 * (y^lambda - 1)
newlm = lm(w ~ X)
```

```
qqnorm(newlm$residuals)
qqline(newlm$residuals)
```



H_0: the errors are normally distributed H_1: the errors are not normally distributed

```
shapiro.test(newlm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newlm$residuals
## W = 0.99734, p-value = 0.5954
```

Since p-value $= 0.5954$, which is bigger than α, we don't have enough to reject H_0.

Fail to reject H_0, so the errors are normally distributed.

```
data3 <- data2 %>% mutate(belowmedian = Height<median(Height)) %>% mutate(w = k1 * (Weight^lambda - 1))
data3_1 <- data3 %>% filter(belowmedian)
data3_2 <- data3 %>% filter(!belowmedian)
glimpse(data3_1)
```

```
## Rows: 252
## Columns: 7
## $ Player      <chr> "Nate\xa0Robinson", "Isaiah\xa0Thomas", "Phil\xa0Pressey",~
## $ Pos         <chr> "G", "G", "G", "G", "G", "G", "G", "G", "G", "G", "G", "G"~
## $ Height      <int> 69, 69, 71, 71, 71, 71, 72, 72, 72, 72, 72, 72, 72, 72, 72~
## $ Weight      <int> 180, 185, 175, 176, 195, 157, 180, 205, 175, 185, 190, 189~
## $ Age         <int> 29, 24, 22, 20, 25, 30, 25, 27, 28, 30, 31, 24, 28, 22, 25~
## $ belowmedian <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE~
## $ w           <dbl> 1961.175, 1967.379, 1954.763, 1956.063, 1979.214, 1929.750~
```

```r
glimpse(data3_2)
```

```
## Rows: 253
## Columns: 7
## $ Player      <chr> "Brandon\xa0Bass", "Reggie\xa0Evans", "Mirza\xa0Teletovic"~
## $ Pos         <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F"~
## $ Height      <int> 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80~
## $ Weight      <int> 250, 245, 235, 230, 260, 220, 218, 248, 230, 235, 240, 230~
## $ Age         <int> 28, 33, 28, 29, 22, 25, 24, 28, 27, 33, 20, 25, 23, 30, 28~
## $ belowmedian <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA~
## $ w           <dbl> 2033.568, 2029.240, 2020.261, 2015.601, 2041.926, 2005.910~
```

```r
n1 <- nrow(data3_1)
n2 <- nrow(data3_2)
lm3_1 <- lm(w ~ Height, data3_1)
lm3_2 <- lm(w ~ Height, data3_2)
di1 <- abs(lm3_1$residuals - median(lm3_1$residuals))
di2 <- abs(lm3_2$residuals - median(lm3_2$residuals))
d1 <- mean(di1)
d2 <- mean(di2)
s1_sq <- var(di1)
s2_sq <- var(di2)
var <- ((n1-1)*s1_sq + (n2-1)*s2_sq) / (n1 + n2 -2)
(d1 - d2) / sqrt(var*(1/n1 + 1/n2))
```

```
## [1] 0.09923723
```

```r
qt(0.975, n1 + n2 -2)
```

```
## [1] 1.964691
```

Since $|t_{BF}| = 0.09923723$, which is smaller than $t_{0.975,n-2} = 1.964691$, we don't have enough evidence to reject H_0. Thus, the error variance is equal.